Towards a logic-based method to infer provenance-aware molecular networks

Zahira Aslaoui-Errafi[1,2], Sarah Cohen-Boulakia[1,2], Christine Froidevaux[1,2⋆], Pauline Gloaguen[3], Anne Poupon[3], Adrien Rougny[1,2], Meriem Yahiaoui[1,2]

[1] Laboratoire de Recherche en Informatique (LRI), CNRS UMR 8623
Université Paris Sud, F-91405 Orsay, Cedex, France
[2] AMIB group, INRIA Saclay
[3] BIOS group, INRA, UMR85, Unité Physiologie de la Reproduction et des Comportements, F-37380 Nouzilly, France ; CNRS, UMR7247, F-37380 Nouzilly, France ; Université François Rabelais, F-37041 Tours, France

Abstract. Providing techniques to automatically infer molecular networks is particularly important to understand complex relationships between biological objects. We present a logic-based method to infer such networks and show how it allows inferring signalling networks from the design of a knowledge base. Provenance of inferred data has been carefully collected, allowing quality evaluation. More precisely, our method (i) takes into account various kinds of biological experiments and their origin; (ii) mimics the scientist's reasoning within a first-order logic setting; (iii) specifies precisely the kind of interaction between the molecules; (iv) provides the user with the provenance of each interaction; (v) automatically builds and draws the inferred network.

1 Context

Biological objects (proteins, genes, small molecules, etc.) interact with each other (physical interactions, activations, inhibitions, catalyses, etc.) and form biological networks. Studying such networks allows the discovery of emerging properties and the understanding of complex biological systems. With the rise of high-throughput experimental methods and the rapid development of bioinformatics analysis methods, there has been a dramatic increase in the quantity and heterogeneity of available data. Providing methods able to automatically construct molecular networks from experimental data is thus of paramount importance. While the series of DREAM workshops competitions have been attempting to address these issues since 2007 (http://www.the-dream-project.org/), they are still challenging. In the ASAM project, we aim to meet the challenging task of designing performant algorithms to infer the topology (or influence graph) of signalling networks (subtask of current DREAM7), a first step in the process of elaborating predictive dynamical models. Given the components of the network,

⋆ Corresponding author: chris@lri.fr. (Authors in alphabetical order). This work has been supported by the INRA-INRIA ASAM project.

building the influence graph consists in determining which relationships they share (e.g., enzymatic catalysis, (de-)phosphorylation, etc.).

Among the approaches available, some provide focused solutions: on connectivity reconstruction (finding which molecules interact with each others from a dataset, without specifying the type of relationship), or on causal reconstruction (finding effects that molecules may have on each others)[20]. Some approaches proceed by refining existing molecular networks [19], which requires the knowledge of an initial network to be analyzed. Others proceed by inferring the network, considering only a single type of experimental method at a time, mostly DNA micro-arrays [1] [4] [3], but also phosphoproteomics [21] [22] [18], or metabolomics [2] etc. Therefore they cannot infer complete pathways involving interactions of different types, failing to provide an holistic view necessary to the understanding of the complexity of living cells [17].

Approaches integrating a variety of different experimental data types have also been proposed [11][24], but they reconstruct only partial network topologies, and define relations without causality (the interactions are not marked as being activations or inhibitions).

Automatic inference of topological networks is performed using various computational approaches that range from statistics [4] to machine learning techniques (e.g. bayesian networks [1]) through logic-based techniques such as boolean networks [25] [12] or fuzzy logic networks [23]. Most of these techniques allow performing model-based prediction as well. In addition, logic-based approaches are particularly well-suited to analyze the consistency of a network towards experimental data by model-checking [9], [7]. Our approach to automatically build the inference graph underlying signalling networks leads to the design of a logic-based model (see Figure 1), in the same spirit as [5] but with a more expressive logical formalism which allows making explicit the expert reasoning. The rest of this paper is organized as follows. Section 2 introduces the Knowledge Base where expert's reasoning is made explicit and a wide variety of experimental data is stored, and shows how the network can be deduced. Section 3 is dedicated to the process of building and drawing the network automatically. Section 4 presents our validation approach while Section 5 draws conclusions and perspectives.

2 Building the Knowledge Base (KB)

Our network inference method is based on the application of reasoning rules to experimental facts. These general rules formalize how experts deduce from experimental facts the belonging of a given molecule to the network and the relationships it shares with other molecules of the network. To achieve the proof-of-concept for this method, we have formalized the knowledge necessary for inferring the signalling network triggered by the FSH receptor [8]. This involved the creation of first-order rules for seven types of experiments.

The design of the KB classically consists of a set of predicates, rules, and facts (ground predicates) described here after.

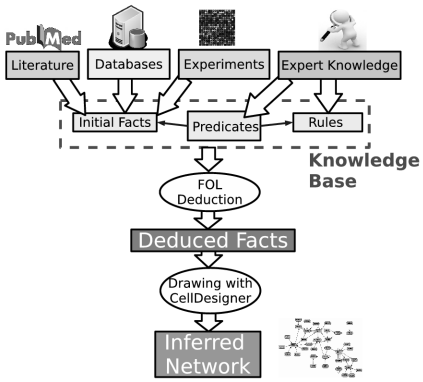

Fig. 1. Workflow of our logic-based method to infer molecular networks

2.1 Predicates

We distinguish three classes of predicates, depending on their biological meaning.
Basic predicates provide ontological types or indicate simple relationships between data (e.g., `AntibodyAgainst(A,X)` means that `A` is an antibody recognizing molecule `X`).
Experiment predicates describe a wide panorama of experiments on molecular data from simple to more complex experiments involving perturbators. An example of a simple experiment is `PA(X,Y,A,E)`, meaning that "in a phosphorylation assay (`PA`) we observe more (resp. less, the same quantity) of the molecule revealed by antibody `A`, which is a phosphorylated form of molecule `Y`, in presence of `X` than in absence of `X`, where `E` can be either 'increase', 'decrease' or 'noeffect'". An example with a perturbator is `ICPPA(X,Y,I,A,E)` meaning that "in a phosphorylation assay (`PA`) in presence of `X` we observe more (resp. less, the same quantity) of the molecule revealed by antibody `A`, which is a phosphorylated form of molecule `Y`, in presence of inhibitor `I` than in absence of `I`, where `E` can be either 'increase', 'decrease' or 'noeffect'". Table 1 shows all the experiment predicates relative to phosphorylation assays (in the variables, the antibody variable `A` is replaced by the molecule variable `Z` it triggers). All the experiments have been formalized by predicates divided into three types depending on the kind of conclusion they allow one to obtain: modulation of a reaction, structure, or localization, and each predicate has been carefully designed and documented as depicted in Table 1.
Network Predicates formalize the different types of biological relationships that can exist between two or more biochemical species (modification, phosphorylation, translation, transcription, complexation, dissociation) taking into account the modulation of these relationships by another species (activation, inhibition or no effect). The modulation of the relationships may be associated with a level of confidence, expressed by the status variable (taking values among: 'invalidated', 'confirmed' or 'hypothesis'). Additionally, the precision of the mod-

Table 1. Phosphorylation assays predicates.

Pred name	Deduction	Method	Variables	Res type	Type (y)	Type (z)	Perturbator (i)
PA, PRA				Phosphoryl.			None
ACPPA	Modulation	Measure	x,y,z,e	of y into z	Protein	Phospho-	Antagonist
ICPPRA	of a	of an	x,y,i,z,e	in the		Protein	Inhibitor
ICPPA	reaction	activity		presence			Inhibitor
SCPPA				of x			siRNA

ulation may vary, which is expressed through the distance variable (taking the values 'direct', 'indirect', or 'unknown'), meaning that a molecule has a direct effect on a relationship (for example, it catalyzes a reaction) or it has an effect via intermediaries, or that this information is currently not available.

2.2 Rules

We have designed two sets of rules. **Simple rules** use experimental data to draw conclusions on new relationships between biochemical species. **Complex** rules use the conclusions obtained by other rules and either provide new data relationships or refine existing relationships. They include transitive rules and rules that deduce an indirect distance (see above) from an unknown distance for a network fact. A simple rule is: `IF PA(X,Y,A,E) AND AntibodyAgainst(A,B) AND PhosphoForm(B,Y,POS) THEN PHOSPHORYLATE(X,Y,B,unknown,confirmed,E)` meaning that "If in a phosphorylation assay where `A` is an antibody that recognizes `B`, `B` is a phosphorylated form of `Y` at position `POS`, we observe more (resp. less, the same quantity) of `B` in presence of `X` than in absence of `X`, then `X` activates (resp. inhibits, does not have any effect on) the phosphorylation of `Y` in `B`, with an unknown distance, where `E` can be 'increase', 'decrease' or 'noeffect'".

2.3 Initial and Deduced Facts

Based on the study of the literature and data from public databases (e.g., SBEAMS, http://www.sbeams.org/), predicates have then been instantiated into facts. Facts have been carefully attached to their **provenance**, specifying the article(s) and/or the data base(s) from which they have been extracted. Various data quality features have been made as explicit as possible in particular the level of confidence to be associated to a set of experimental results.

Initial facts have then been used to produce deduced facts by triggering rules. As expert rules are mostly Horn rules, we have mainly considered using forward chaining procedures for rules-base systems [6], leading to a saturation of the KB. In order to tackle more expressive rules (e.g. typing rules) and facts involving functional terms that describe molecular complexes, we have also been considering first-order consequence-finding techniques. In these settings, provenance information of deduced facts is obtained from their proof(s) making it possible to assess their quality from the quality of initial facts used in the deduction process.

3 Automatic generation of the signalling network

The network topology is automatically built from our initial and deduced facts using the software CellDesigner [15], that is widely used to model molecular networks. CellDesigner uses (and extends) two standards, SBML (Systems Biology Markup Language) [10] and SBGN (Systems Biology Graphical Notation) [16], for the description and representation of molecular networks. CellDesigner can represent: cell compartments, biochemical species (nodes of the network) and reactions (arcs). Each biochemical species is localized in one or more compartment, and assigned to a biological type (e.g., complex, protein). Each reaction has a type (e.g., translation, transcription) and at least one reactant species and one product species, and can be modulated (e.g., activated/catalyzed) by one or more modifier species. The various species, reactions and modulation types are represented in the network by different forms of nodes and arcs (e.g. ovals for simple molecules; plain lines for state transitions).

Our facts (initial and deduced) are automatically translated into the extended version of SBML that CellDesigner uses. Initial facts are used to describe the chemical species of the networks (name, type and residue modifications, e.g a phosphorylation on the residue S132). The deduced facts are used to describe the reactions (type, reactants, products and modulations) and species localizations. The status and the distance of the reaction modulations are modelled by different line-styles of arcs linking a modifier species to a reaction (e.g. black line for a confirmed status or red line for a hypothesis status) in the network. The layout of the network is then automatically calculated by CellDesigner. This automatic process is illustrated by a short example.

Example. Let `phosphorylate(fsh,mek1,pmek1s17,indirect,confirmed,increase)` be a deduced fact to be represented with CellDesigner. First, the fact has to be translated into (extended) SBML. It describes a relation between the molecules (first three arguments) `fsh`, `mek1` and `pmek1s17` (the latter being phosphorylated-mek1 at position S17). To declare precisely these molecules into CellDesigner, the KB is queried to get their type (`protein(fsh)`, `protein(mek1)` and `phosphoprotein(pmek1s17))`, their localization in the cell (`localization(fsh,extracellular)`, `localization(mek1,cytosol)` and `localization(pmek1s17,cytosol)`) and their possible residue modifications (`phosphoform(pmek1s17,mek1,s17)`). This is given by initial facts of the KB that are retrieved by exploiting the predicates typology. Second, the deduced fact is translated into an SBML reaction. Here, the type of reaction is a state transition (`phosphorylation`), the reactant is the second argument (`mek1`), the product the third (`pmek1s17`) and the reaction is activated (`increase`) by the first argument (`fsh`). Because the modulation of the reaction has a status `confirmed` and is `indirect`, the line-style of the arc representing the modulation is declared as being black and dashed.

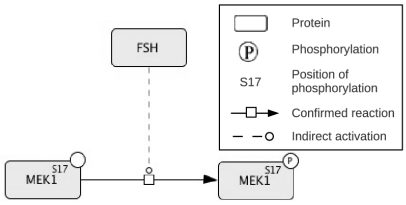

Fig. 2. Building elements of network with CellDesigner

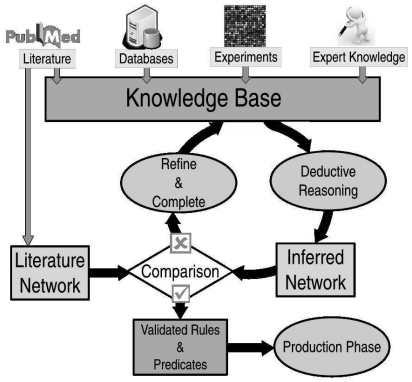

Fig. 3. Workflow for validation of our method to infer molecular networks

4 Validation

A phase of trial and error may be necessary in order to validate the method, before using it in production phase. This validation phase will require an important accumulation of data. Indeed, we need to infer automatically the networks from experimental data and compare them to networks of the literature to seek for missing links (links absent while expected), indicating missing reasoning expert rules which will need to be written. This upstream work has already been done for the follicle-stimulating hormone receptor (FSH-R) [8]. Even with this quite simple network, over 400 initial facts and 100 expert rules were necessary. Once processed, the KB contained over 1,000 (initial and deduced) facts. Two different serotonin receptors (5HT2 and 5HT4) and Epidermal Growth Factor receptor (EGF-R) are currently considered. We envisage to speed up the detection of biological objects in the literature that will be used within facts by using text mining tools such as PathText [14]. We expect that after studying a few networks, expert rules will be complete, validating the KB. During this validation phase and the following production phase, the inferred networks will be analyzed and all the unexpected conclusions will be validated experimentally.

5 Conclusion and Discussion

We have introduced a bottom-up method to infer molecular networks whose originality lies in its knowledge-driven basis. More precisely: (i) domain knowledge is collected from experts, literature and public databases, and data from classical and high-throughput experiments; (ii) the scientist's reasoning is mimicked within an expressive logical setting (first-order logic); (iii) the kind of interaction between the molecules is precisely specified; (iv) provenance of deduced facts (translated into interactions) is provided, allowing users to evaluate the quality of the inferred pieces of knowledge; (v) the inferred network is automatically built and drawn, offering the possibility of interacting with it.

Such an automated method, allowing automated exploitation of experimental data and resulting in molecular networks, will become increasingly necessary with the spread of huge amounts of data produced by large-scale experimental methods.

Ongoing work include considering techniques to formalize efficiently expressive rules and facts (including considering encoding into a propositional setting). Importantly, while all our current efforts have been made on deductive reasoning, we need also to consider exploring abductive reasoning to help scientists understanding how the biological system can be perturbed to reach a desired behaviour. Users may wish to know which facts or rules are to be added to reinforce the reliability of an existing arc or to add a new arc in the network. Determining the rules and pieces of data to be added to the initial KB to get the desired behaviour is obtained by the abduction reasoning that will provide all the possible hypotheses expressed in terms of initial facts and logical rules. Advices will thus be given to the scientists about the new experiments to be conducted to validate some hypothesis. As a consequence, we are currently looking at systems able to provide both deductive and abductive reasoning and are considering the SOLAR system as a first excellent candidate [13].

References

1. M. Bansal, V. Belcastro, A. Ambesi-Impiombato and D. di Bernardo. *How to infer gene networks from expression profiles*, Mol Syst Biol., 3:78, 2007.
2. T. Cakir, M. Hendriks, J.A. Westerhuis and A.K. Smilde, *Metabolic network discovery through reverse engineering of metabolome data.*, Metabolomics, 5:318-29, 2009.
3. J. Chiquet, A. Smith, G. Grasseau, C. Matias and C. Ambroise, *SIMoNe: Statistical Inference for MOdular NEtworks.*, Bioinformatics, 25:417-8, 2009.
4. I. Drozdov, B. Svejda, B.I. Gustafsson, S. Mane, R. Pfragner, M. Kidd and I.M. Modlini, *Gene Network Inference and Biochemical Assessment Delineates GPCR Pathways and CREB Targets in Small Intestinal Neuroendocrine Neoplasia.*, PLoS ONE, 6(8), 2011.
5. F. Eduati, A. Corradin, B. Di Camillo and G. Toffolo, *A Boolean approach to linear prediction for signaling network modeling*, PLoS ONE, 5(9), 2010.
6. C. Forgy, *Rete: A Fast Algorithm for the Many Pattern-Many Object Pattern Match Problem*, Artificial Intelligence, 19:17-37, 1982.

7. M. Gebser, T. Schaub, S. Thiele and Ph. Veber, *Detecting inconsistencies in large biological networks with answer set programming*. Theory and Practice of Logic Programming, 11(2-3):323-360, 2011.
8. P. Gloaguen, P. Crépieux, D. Heitzler, A. Poupon and E. Reiter, *Mapping the follicle-stimulating hormone-induced signaling networks.*, Front Endocrinol., 2:45, 2011.
9. C. Guziolowski, J. Gruel, O. Radulescu and A. Siegel, *Curating a Large-Scale Regulatory Network by Evaluating Its Consistency with Expression Datasets*, CIBB, 144-155, 2008.
10. M. Hucka *et al.*, *The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.*, Bioinformatics, 19(4):524-31, 2003.
11. D. Hwang, J.J. Smith, *et al.*, *A data integration methodology for systems biology: Experimental verification*, PNAS, 102(48):17302-7, 2005.
12. K. Inoue, *Logic Programming for Boolean Networks.*, Proc. of IJCAI, 924-930, 2011.
13. K. Iwanuma, K. Inoue, O. Ray, *SOLAR: An automated deduction system for consequence finding*. AI Commun. 23(2-3): 183-203, 2010.
14. B. Kemper, T. Matsuzaki, *et al. PathText: a text mining integrator for biological pathway visualizations*. Bioinformatics [ISMB]. 26(12):374-381, 2010.
15. H. Kitano, A. Funahashi, Y. Matsuoka and K. Oda, *Using process diagrams for the graphical representation of biological networks.*, Nat Biotechnol., 23(8):961-6, 2005.
16. N. Le Novère *et al.*, *The Systems Biology Graphical Notation.*, Nat Biotechnol., 27(8):735-41, 2009.
17. F. Markowetz and R. Spang, *Inferring cellular networks - a review*, BMC Bioinformatics, 8 Suppl 6:S5, 2007.
18. A. Mitsos, I.N. Melas, P. Siminelakis, A.D. Chairakaki, J. Saez-Rodriguez and L.G. Alexopoulos, *Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data.*, PLoS Comput Biol., 5(12), 2009.
19. M.K. Morris, J. Saez-Rodriguez, D.C. Clarke, P.K. Sorger and D.A. Lauffenburger, *Training Signaling Pathway Maps to Biochemical Data with Constrained Fuzzy Logic: Quantitative Analysis of Liver Cell Responses to Inflammatory Stimuli*, PLoS Comput Biol., 7(3), 2011.
20. N. Papin, H. Tony, O.P. Bernhard and S. Shankar, *Reconstrution of cellular signalling networks and analysis of their properties*, Nat Rev Mol Cell Biol., 6(2):99-111, 2005.
21. R.J. Prill, J. Saez-Rodriguez, L.G. Alexopoulos, P.K. Sorger and G. Stolovitzky, *Crowdsourcing Network Inference: The DREAM Predictive Signaling Network Challenge.*, Sci Signal, 4:mr7, 2011.
22. K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger and G.P. Nolan, *Causal protein-signaling networks derived from multiparameter single-cell data.*, Science, 308(5721):523-9, 2005.
23. J. Saez-Rodriguez, L.G. Alexopoulos, J. Epperlein, R. Samaga, D.A. Lauffenburger, S. Klamt and P.K. Sorger. *Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction.*, Mol Syst Biol. 5:331, 2009.
24. C.H. Yeang, T. Ideker and T. Jaakkola, *Physical Network Models*, J Comput Biol., 11(2-3):243-62, 2004.
25. D. Wittmann, J. Krumsiek, J. Saez-Rodriguez, D.A. Lauffenburger, S. Klamt and F. Theis, *From Qualitative to Quantitative Modeling*, BMC Syst. Biol. 3:98, 2009.